Poster II-54

Enhancing Access to the Bibliome: The TREC Genomics Track

Hersh, William
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

The growing amount of scientific discovery in genomics and related biomedical disciplines has led to a corresponding growth in the amount of on-line data and information. A growing challenge for biomedical researchers is how to access and manage this ever-increasing quantity of information. This situation presents opportunities and challenges for the information retrieval (IR) field. The Text Retrieval Conference (TREC) has implemented a Genomics Track to create an experimental environment for work at the interface between biomedical and IR research.

TREC is an annual activity of the IR community aiming to evaluate systems and users. It is sponsored by the National Institute for Standards and Technology (NIST, trec.nist.gov). IR has historically focused on document retrieval, but the field has expanded in recent years with the growth of new information needs (e.g., question-answering, cross-lingual), data types (e.g., video) and platforms (e.g., the Web). TREC activity is organized into "tracks" of common interest, such as question-answering, multi-lingual IR, Web searching, and interactive retrieval. TREC generally works on an annual cycle, with data distributed in the spring, experiments run in the summer, and the results presented at the annual conference that usually takes place in November. A new Genomics Track was established in 2003.

The track was organized around two tasks: an ad hoc retrieval task (i.e., conventional searching) and an information extraction task. The ad hoc retrieval task was guided by the availability of Gene Reference into Function (GeneRIF) data in the LocusLink database. Each GeneRIF entry consists of a statement about the function of a gene along with a pointer to the MEDLINE reference for the article that discovered that data. We randomly selected 50 genes to use as topics from those having GeneRIFs. The document collection for searching was a one-year subset of MEDLINE. Researchers were allowed to use any external resource for their experiments with the exception of the GeneRIF field of LocusLink. A total of 26 research groups submitted results.

The information extraction task was an exploratory task of extracting the GeneRIF statement from the MEDLINE record or the article proper. Research groups were provided both, with lexical overlap of the GeneRIF statement as measured by the Dice coefficient and some variations of it serve as the measures for success. Full-text articles were provided through Highwire Press (www.highwire.org).

Future years of the track will include the addition of other information resources and user experiments.

Grant Support: NSF ITR grant 0325160.